A simulation framework of high-dimensional phylogenetic microbiota data

Perrine Soret[1,2,3,4*], Marta Avalos[1,2,3**], Laurence Delhaes[2,5,6] and Rodolphe Thiébaut[1,2,3,4,7]

[1]INRIA, SISTM team, Bordeaux, France
[2]University of Bordeaux, Bordeaux, France
[3]Bordeaux Population Health Center INSERM U1219, Bordeaux, France
[4]Vaccine Research Institute, Créteil, France.
[5]Cardiothoracic Research Center, INSERM U1045, Bordeaux, France
[6]CHU Bordeaux, Department of Parasitology - Mycology, Bordeaux, France
[7]CHU Bordeaux, Department of Public Health, Bordeaux, France
*Presenting author: *perrine.soret@u-bordeaux.fr* **Corresponding author: *marta.avalos-fernandez@u-bordeaux.fr*

Context

The increasing quality/reducing cost of high-throughput sequencing technology, in particular, 16S rRNA gene sequencing of the bacterial component (and to a lesser extent, ITS2 sequencing of the fungal component) of the human microbial community (microbiota), has enabled researchers to investigate human diseases. Subsequently, microbiota has been associated with numerous diseases, including inflammatory bowel disease, diabetes, cancer and cystic fibrosis.
The microbiota sequencing data are measured as reads' counts (often with an excess of zeros), interpreted as a taxon's abundance in a microbial community. To make the microbial abundance comparable across samples, data are typically normalized to the relative abundances of all bacteria observed, that is an example of the so-called Compositional Data (CoDa).
CoDa consists of a collection of nonnegative measurements that sum to a constant value, typically, proportions that sum to 1. Because knowing the sum, one component can be determined from the sum of the remainder, the parts that make up the composition are mathematically and statistically dependent. In general, CoDa are mapped from the constrained simplex space to the Euclidian space using nonlinear transforms to allow valid inferences. Also, the microbiota data are organized under a phylogenetic structure that, if deeply assessed, lead to high-dimensionality. In parallel, in response to the needs, there is an intensive emergence of specific statistical methods and computational tools. Because of the recentness, it is still too soon to evaluate the applicability and accuracy of available methods.
Simulation studies, in which a sample of random data is computationally generated many times mimicking a real data distribution, are a standard tool to compare the performance of competitive statistical methods.

Notations

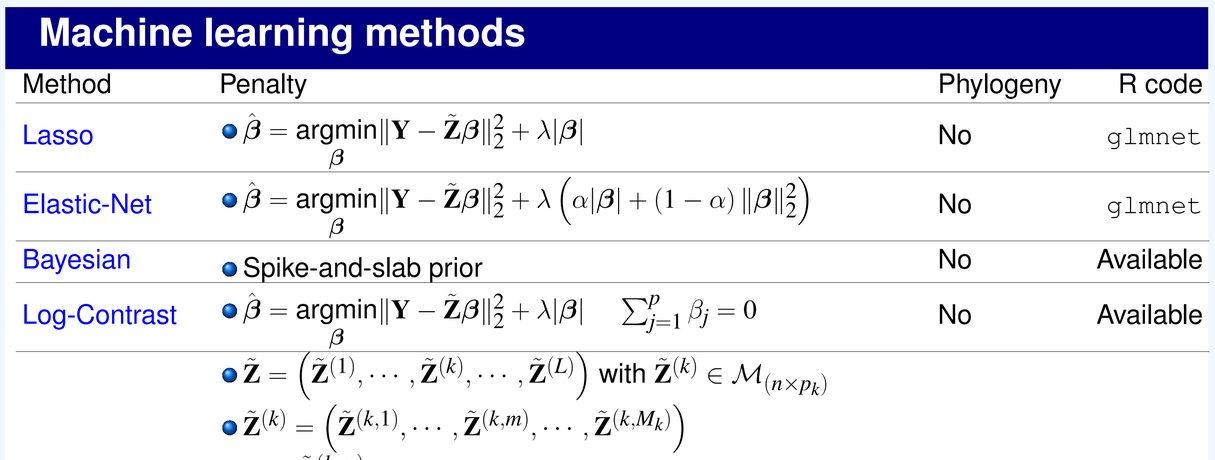
- n number of subjects ; q number of sample OTUs ; p number of covariates
- $Z_{ij} \in \mathbb{N}^+$ raw abundance ; $\tilde{Z}_{ij} = \frac{Z_{ij}}{\sum_{j=1}^{q} Z_{ij}} \in [0,1]$ relative abundance of patient i for OTU j
- $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)^\top$ and $\tilde{\mathbf{Z}} = \left(\tilde{\mathbf{Z}}_1, \cdots, \tilde{\mathbf{Z}}_n\right)^\top$ absolute and relative abundance matrices $n \times q$.
- $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^\top$ matrix of covariates $n \times p$; $\mathbf{Y} = (Y_1, \cdots, Y_n)$ outcome vector $n \times 1$

Machine learning methods

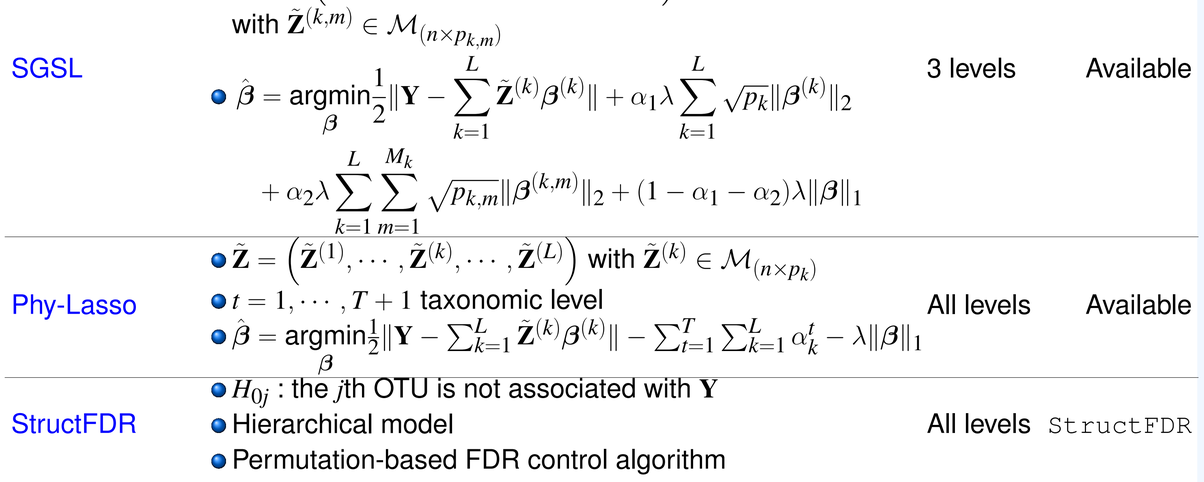
Method	Penalty	Phylogeny	R code
Lasso	$\hat{\beta} = \underset{\beta}{\text{argmin}} \|\mathbf{Y} - \tilde{\mathbf{Z}}\beta\|_2^2 + \lambda\|\beta\|$	No	`glmnet`
Elastic-Net	$\hat{\beta} = \underset{\beta}{\text{argmin}} \|\mathbf{Y} - \tilde{\mathbf{Z}}\beta\|_2^2 + \lambda\left(\alpha\|\beta\| + (1-\alpha)\|\beta\|_2^2\right)$	No	`glmnet`
Bayesian	Spike-and-slab prior	No	Available
Log-Contrast	$\hat{\beta} = \underset{\beta}{\text{argmin}} \|\mathbf{Y} - \tilde{\mathbf{Z}}\beta\|_2^2 + \lambda\|\beta\|$ $\quad \sum_{j=1}^{p} \beta_j = 0$	No	Available
SGSL	$\tilde{\mathbf{Z}} = \left(\tilde{\mathbf{Z}}^{(1)}, \cdots, \tilde{\mathbf{Z}}^{(k)}, \cdots, \tilde{\mathbf{Z}}^{(L)}\right)$ with $\tilde{\mathbf{Z}}^{(k)} \in \mathcal{M}_{(n \times p_k)}$; $\tilde{\mathbf{Z}}^{(k)} = \left(\tilde{\mathbf{Z}}^{(k,1)}, \cdots, \tilde{\mathbf{Z}}^{(k,m)}, \cdots, \tilde{\mathbf{Z}}^{(k,M_k)}\right)$ with $\tilde{\mathbf{Z}}^{(k,m)} \in \mathcal{M}_{(n \times p_{k,m})}$; $\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2}\|\mathbf{Y} - \sum_{k=1}^{L} \tilde{\mathbf{Z}}^{(k)}\beta^{(k)}\| + \alpha_1\lambda\sum_{k=1}^{L}\sqrt{p_k}\|\beta^{(k)}\|_2 + \alpha_2\lambda\sum_{k=1}^{L}\sum_{m=1}^{M_k}\sqrt{p_{k,m}}\|\beta^{(k,m)}\|_2 + (1-\alpha_1-\alpha_2)\lambda\|\beta\|_1$	3 levels	Available
Phy-Lasso	$\tilde{\mathbf{Z}} = \left(\tilde{\mathbf{Z}}^{(1)}, \cdots, \tilde{\mathbf{Z}}^{(k)}, \cdots, \tilde{\mathbf{Z}}^{(L)}\right)$ with $\tilde{\mathbf{Z}}^{(k)} \in \mathcal{M}_{(n \times p_k)}$; $t = 1, \cdots, T+1$ taxonomic level ; $\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2}\|\mathbf{Y} - \sum_{k=1}^{L} \tilde{\mathbf{Z}}^{(k)}\beta^{(k)}\| - \sum_{t=1}^{T}\sum_{k=1}^{L}\alpha_k^t - \lambda\|\beta\|_1$	All levels	Available
StructFDR	H_{0j} : the jth OTU is not associated with $\mathbf{Y}$; Hierarchical model ; Permutation-based FDR control algorithm	All levels	`StructFDR`

Results

Protocol Simulation S1



Protocol Simulation S2



CoDa

Absolute microbiome abundances are not informative

Raw abundance
$$\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iq})^\top \in \mathbb{R}_+^q$$

Relative abundance
$$\tilde{\mathbf{Z}}_i = (\tilde{Z}_{i1}, \cdots, \tilde{Z}_{iq})^\top \in \mathbb{S}^q$$

$$\mathcal{C}(\mathbf{Z}_i) = \left(\frac{Z_{i1}}{\sum_{j=1}^{q} Z_{ij}}, \cdots, \frac{Z_{iq}}{\sum_{j=1}^{q} Z_{ij}}\right) = (\tilde{Z}_{i1}, \cdots, \tilde{Z}_{iq})^\top = \tilde{\mathbf{Z}}_i \qquad i = 1, \cdots, n$$

Simplex : $\mathcal{S}^q = \left\{\tilde{\mathbf{Z}}_i = (\tilde{Z}_{i1}, \cdots, \tilde{Z}_{iq})^\top : \ \tilde{Z}_{ij} > 0, \ j = 1, \cdots, q; \ \sum_{j=1}^{q} \tilde{Z}_{ij} = 1\right\}$

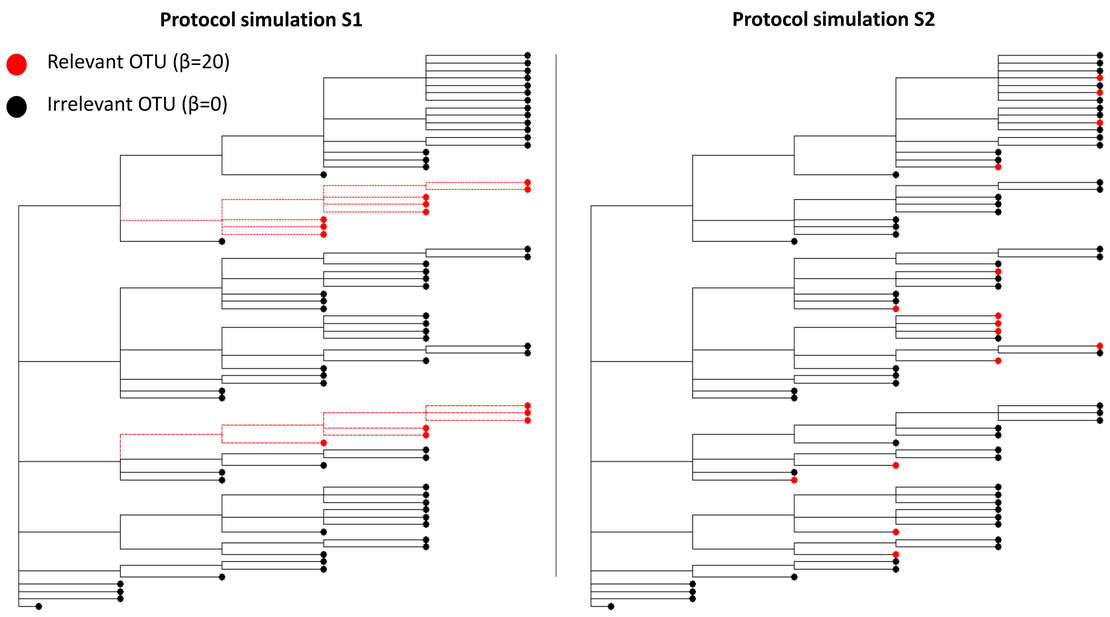
Ignoring the compositional nature of the data may induce strong incoherences in correlations and distances

Conditions to apply statistical methods
- Permutation-invariant
- Scale invariance
- Subcompositional coherence

(Main) Log-ration transformations
- Additive log-ratio (alr)
- Centered log-ratio (clr)
- Isometric log-ratio (ilr)

Simulation Protocols

- π estimated on real data (mean of OTUs)
- $\tilde{\mathbf{Z}} \sim \text{DM}(\pi) \in \mathcal{M}_{n \times p}$ with $n = 30, 100$ and $p = 75$
- β choosen following two protocols
- $\mathbf{Y} = 54 + \text{alr}\left(\tilde{\mathbf{Z}}\right)\beta + \varepsilon$
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- σ^2 determined by signal-to-noise ratio



Tuning parameter selection and comparison criteria

Cross-Validation : 10 fold-CV such as $CV(\lambda) = \frac{1}{n}\sum_{i=1}^{K}\frac{1}{n_k}\sum_{i \in \mathbf{D}_k}\left(Y_i - \mathbf{Z}_i\hat{\beta}(\lambda)_{\mathbf{D}_{\setminus k}}\right)^2$

Bootstrap : Select OTUs stably associated to the outcome (100 bootstrap samples)

Selection proportion (SP) : $SP\left(\hat{\beta}_j\right) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}_{\hat{\beta}_{j,s} \neq 0} \quad \in \quad [0,1]$

Conclusion

The complexity of microbiome data makes generating realistic data challenging. The negative false rate represents the rate of relevant variable not selected and, conversely, the positive false rate represents the rate of irrelevant variable selected by the method. For the two protocols, the simplest methods (Lasso, ElasticNet) showed better results than the most complex methods (as Log-Contrast or Phylasso). SGSL and the bayesian method selected a higher number of predictors. When the relevant variables are randomly generated (protocol S2), the selection appeared to be more instable : the rate of false negative presented higher variability between replications.
Globally, when the generated data are relatively simple, as expected, the simplest methods show a better behavior. The most complex methods may outperform in most complex situations (for example, when several OTUs are correlated or when the phylogenetic structure influences the outcome). However, the impact of generating correlated OTUs or accounting for the phylogenetic structure is unclear (Liu, 2015, Rush, 2016).
The R Code for the simulation framework is available on
https ://github.com/psBiostat/SimulationMicrobiomeData.git

References

- Tanya P Garcia, Samuel Müller, Raymond J Carroll, and Rosemary L Walzem, *Identification of important regressor groups, subgroups and individuals via regularization methods : application to gut microbiome data*, Bioinformatics **30** (2014), no. 6, 831–837.
- Zhenqiu Liu, Shili Lin, and Steven Piantadosi, *Network construction and structure detection with metagenomic count data*, BioData mining **8** (2015), no. 1, 40.
- Stephen T Rush, Christine H Lee, Washington Mio, and Peter T Kim, *The phylogenetic lasso and the microbiome*, arXiv preprint arXiv :1607.08877 (2016).
- Jyoti Shankar, Sebastian Szpakowski, Norma V Solis, Stephanie Mounaud, Hong Liu, Liliana Losada, William C Nierman, and Scott G Filler, *A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses*, BMC bioinformatics **16** (2015), no. 1, 31.
- Jian Xiao, Hongyuan Cao, and Jun Chen, *False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing*, Bioinformatics **33** (2017), no. 18, 2873–2881.